DOCUMENT RESUME

ED 135 836 TM 006 061

AUTHOR Sachar, Jane
TITLE Use of Performance Models in Developing Standardized Tests.
PUB DATE [Sep 76]
NOTE 15p.; Paper presented at the Annual Meeting of the American Psychological Association (84th, Washington, D.C., September 3-7, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Arithmetic; Elementary Education; *Elementary School Students; Grade Equivalent Scores; Item Sampling; Norms; Probability; *Standardized Tests; *Test Construction
IDENTIFIERS Performance Models

ABSTRACT

A test of mental arithmetic skills was developed and normed for children in grades one to six. Statistical techniques were used to predict the probability correct for each item and to rank order the items. Students in a given grade took only one level of the test and inferences were made of their performance on those tests which they did not take. Using performance models, student scores on each test were predicted by estimating parameters in functional relationships between probability correct for the population and the individual. A function was found mapping the probability correct for a given item on one test and grade to the probability correct for the same item on another test and grade. Then the probability correct on any item in any test could be predicted for a given grade. Performance models for each student were used to predict student scores on each test. A functional relationship was found between the probability correct for the population and the probability correct for the individual on groups of items. From the student's performance curves, the probability correct for those items that he or she did not take could be predicted, and a total score on any given test thus computed. Traditional norming procedures were then applied to these scores in determining grade equivalents. (Author/RC)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

ED135836

# USE OF PERFORMANCE MODELS IN DEVELOPING STANDARDIZED TESTS

Jane Sachar

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Navy Personnel Research and Development Center
San Diego, California 92152

Paper presented at the annual meetings of the
American Psychological Association, Washington, D. C.
September 3-7, 1976

TM006 061

This paper discusses the development and norming of six levels of a test of mental arithmetic skills designed for grades 1-6. Mental arithmetic skills, which are distinguished from regular arithmetic skills, include no supporting visual stimulus. Each test consists of a set of 100 or 110 exercises, presented using a prerecorded tape, allowing the student 6 seconds to answer each item. Students respond on prepared answer sheets and written free-response answers are expected. Students are instructed to write only answers, not problem statements.

Through the years, proficiency in mental arithmetic lost its emphasis in the elementary mathematics curriculum, giving way to new educational philosophies. Recently, however, educators have again recognized the importance of proficiency in mental arithmetic, and it is becoming a more widely-taught skill.

In reviewing the literature on mental arithmetic, we found few quantitative studies and no standardized tests. In recognition of the need for a test that assesses mental arithmetic skills, we have undertaken to develop one.

The first step in the development of the test was to define the mental arithmetic curriculum, or the universe of items. The curriculum used in the present work was adapted from the drill-and-practice arithmetic strands program developed at the Institute for Mathematical Studies in the Social Sciences (IMSSS) at Stanford University (Suppes, Jerman, & Brian, 1968; Suppes & Morningstar, 1972; Suppes, Searle, & Lorton, note 1). The curriculum covers a portion of the standard mathematics curriculum that can be presented orally, including seven concepts, or strands; namely, addition, subtraction, multiplication, division, number concepts, measurement, and fractions. The exercises of a single type within a strand constitute

an equivalence class. Each equivalence class has an explicit definition in terms of the format of the exercise, the operation used, the size of the numbers involved, and so on.

The items to be included in each test were then selected using a stratified item sampling scheme. The test at each grade level covers items with probability correct (P) ranging from 0 to 1, thus avoiding a floor or ceiling effect. The exercises are ordered by difficulty levels (as determined by pretesting), with the easier ones appearing in the first part of the test. Thus, a student's grade equivalence score roughly corresponds to his "quitting" point. We found that, with this test structure, students are more likely to give their full attention and effort to exercises that they are able to do. It is unlikely that an item that a student can do will be embedded in a group of items he cannot.

Construction of the primary tests was carried out as follows (see Olshen, 1975, for details):

1. A set of equivalence classes was identified that contained material appropriate for each of the grades. Three of the seven strands were represented in Grades 1 through 3, whereas all seven were represented in Grades 4-6.

2. Performance data were obtained for exercises sampled from these classes, and the data were used to establish an ordering of classes, from least difficult to most difficult.

3. For each grade level, an interval along the resulting continuum of ordered classes was chosen with the proper boundary conditions, that is, P close to 1 at the left and to 0 at the right. For example, an easy first grade item is "What is 2+3?" and a difficult sixth grade item is "What minus 33 equals 25?" The resulting intervals have large regions

of overlap between grades.

4. For each grade, the universe of equivalence classes in the interval was divided into several strata, further dividing the four basic operations by format.

5. A family of tests for each grade level was defined by the number of equivalence classes from each stratum, as well as the range of classes. With this constraint on the distribution of classes, the particular equivalence classes from each of the strata were chosen at random from the set of all possible equivalence classes in that stratum for that grade.

6. A test was then constructed by randomly generating one exercise from each of the selected equivalence classes.

For our purposes two parallel test forms were constructed using this procedure.

The norming population involves determining "grade equivalent" scores for each grade on each level of the test. Such grade equivalents are derived by finding the median student score; that is, the score of the student above which 50% of the student's scores fall. It was not possible either to give all the tests to students at all grade levels or to spread administration of the tests across a school year. For this reason techniques were developed to make inferences of the performance of students on tests that they did not take.

To make such inferences, I need two functions relating probabilities correct. Consider student S in Grade G. The first function predicts the probability correct on an item for students in G from the performance on the same item by students in another grade. The second function predicts the probability correct on an item for S from the probability

correct on the same item for all students in G. With these two functions, I can predict the probability correct for S on each item he did not take, which, in turn, yields an estimate of S's score on a test he did not take.

The first function to be found was that mapping the probability correct for an item on test x to the probability correct for the same item on test y. Because no two tests contained identical items, a function mapping equivalence classes rather than items was used. The two tests were taken by two different grades. The overlap of ranges of equivalence classes for pairs of tests was quite large, producing many equivalence classes covered in the ranges of all tests. The desired function was determined for each pair of grades using the proportion correct on all equivalence classes responded to by both grades. If an equivalence class appeared on both forms for a given grade, the probability correct was taken to be an average of the two. Using these data, the function and the values of its parameters were determined. Let t(i) = 1 if an item from equivalence class i on test x was answered correctly and 0 if it was answered incorrectly. Then the proportion of students answering the item correctly on test x is a function, F, of the proportion of students answering the item correctly on test y.

$$P[\, t(i) \text{ on } x = 1 \,] = F\,(\, P[\, t(i) \text{ on } y = 1 \,]\,)$$

The model fitting the data was

$$\arcsin\,(y(i)^{1/2}) = a + b * \arcsin\,(x(i)^{1/2}),$$

where y(i) = probability correct for one grade on items from equivalence class i, and x(i) = probability for another grade on items from equivalence class i.

The arcsin model is based on a variance stabilizing transformation on proportions, given by Anscombe (1948), $\arcsin\,(p'^{\,1/2})$ where

$p' = (r + 3/8)/(n + 3/4)$,

r = the number of students answering the item correctly, and

n = the total number of students taking the test.

The tails of the distribution of $p = r/n$ are pushed toward the center, where $p = .5$, to produce the distribution of $p'$. Figures 1 and 2 display examples of a strong relationship when the number of overlapping items is large (99) and small (30), respectively. The two curves shown are the fits of y on x and x on y.

The second function to be found relates performance by the individual to performance by the grade on an item. Each student has his own graph. The probability correct for the grade determines the location on the abscissa for a given item, the abscissa being P [ t(i) = 1 ], written p(i). On the ordinate falls the observed probability correct by a given student.

Because each item in this study has a binary response variable with a different probability of success, the items were grouped (Cox, 1970). Based on the assumption that for a given student the probability of success, p(k)*, is constant within set k, the data were grouped into sets disregarding the individual responses. Let there be m items in each set, R(k) of which are responded to successfully. Then

$E( R(k)/m ) = p(k)^*$

$Var( R(k)/m ) = p(k)^* ( 1 - p(k)^*)/ m.$

The sets are formed using the proportion correct, p(i), on the item by the grade. Let the 220 items (110 items from each of two test forms) be ordered such that pi[1] < pi[2] <...<pi[220]. The first set consists of items i[1], ... , i[m]. The j-th set consists of items i[m(j-1)+1], ... , i[(m)(j)]. We formed sets of m = 10 items, resulting in approxi-

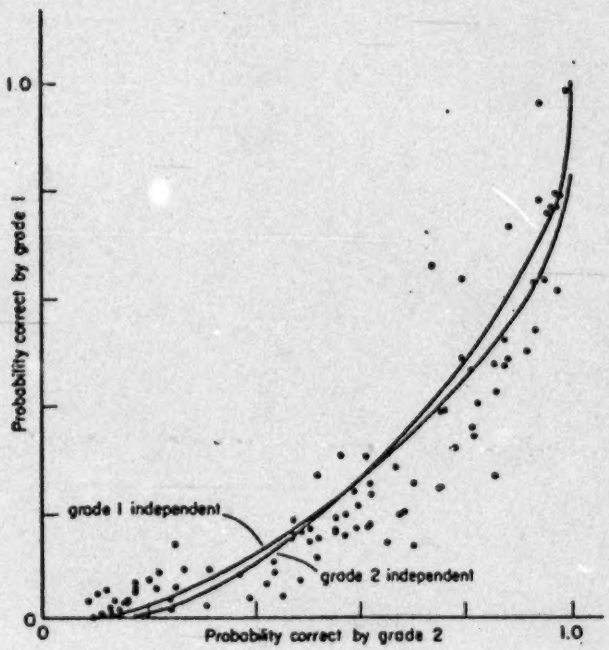


Figure 1. Probability correct for Grades 1 and 2 (N=99) on identical equivalence classes.

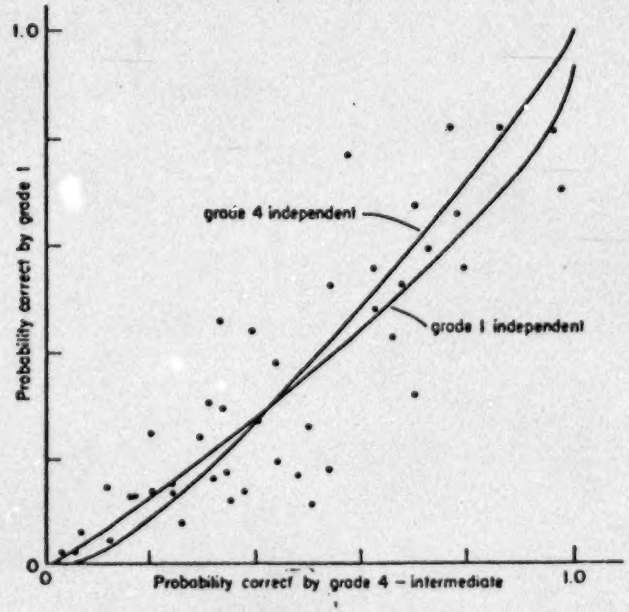


Figure 2. Probability correct for Grades 1 and 4 Intermediate (N=30) on identical equivalence classes.

mately 20 data points for the student.

The points for each student, j, whose score was being predicted were graphed. The performance curves are monotonic increasing curves with the probability correct for all students on the abscissa and the probability correct for the individual student on the ordinate. Presumably the student cannot answer items above a certain difficulty level, can easily answer correctly those below a certain difficulty level, and can answer with a positive probability items between those levels. The performance curves, therefore, have three distinct intervals. The first interval contains only items which the student missed; the second includes both items the student answered correctly and items which he missed; and the third contains only items which the student answered correctly. From this graph, the function $P\ [t(i)=1 \mid j]$ was determined and from this function the score for the student was predicted. Figures 3 and 4 display these graphs for two students.

Five models that describe the student's performance were tested. The best fitting model was $f(x) = k_1 + k_2 x^c$, relating the probability correct for the student, $f(x)$, to the probability correct for the class, $x$. This function produces a linear, concave, or convex curve depending on the value of c.

Student j's score on test k can be expressed as

$$y(j,k) = a(j,k)*p(j,1) + \sum f(p(i)) + c(j,k)*p(j,3)$$

where i refers to item i and

y(j,k) = student j's score on test k,

a(j,k) = number of exercises on test k that are in the initial range where student j got them all correct,

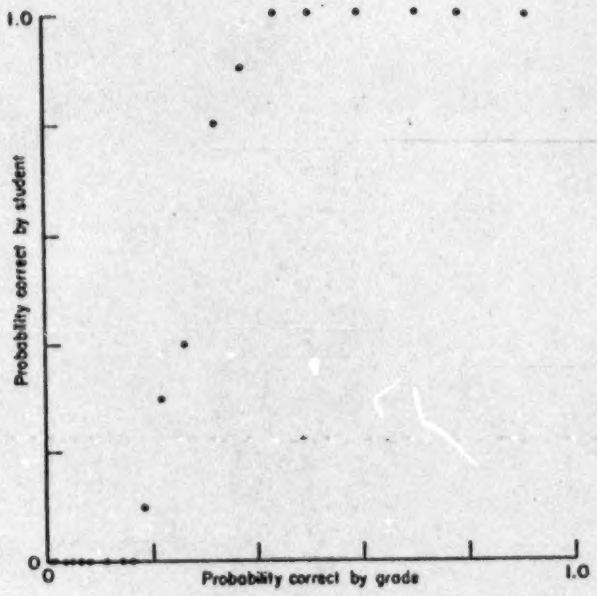


Figure 3. Student Graph No. 1 of probability correct.

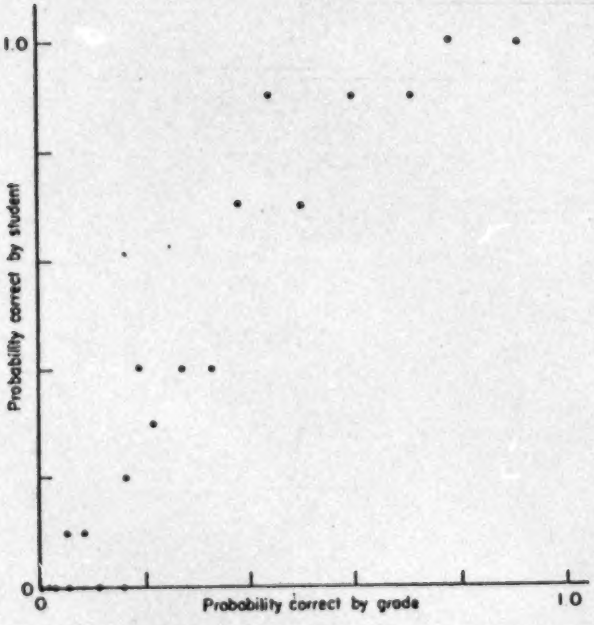


Figure 4. Student Graph No. 2 or probability correct.

f(p(i)) = the function determined below, of exercise i where exercise i appears on test k,

c(j,k) = number of exercises on test k that are in the last range where studnet j got them all wrong,

p(j,1) = probability that student j got the exercises in the first range correct = 1, and

p(j,3) = probability that student j got the exercises in the last range correct = 0.

The student's score then becomes

$$y(j,k) = a(j,k) + \sum f(P(i)).$$

Following the estimation of the parameters, the student's score for a test that he did not take is computed as follows:

Let f.12(i) = predicted proportion correct on item i for Grade 1 on an item that Grade 2 took using the above arcsin regression.

Let g(x) = probability correct for this student on an item with probability correct, x, for Grade 1 (found from the curve fitted for the student).

Then this first-grade student's predicted total score on a Grade 2 test is:

$$T = \sum g(\ f.12(i)\ )/2 \qquad i = 1,\ldots,n$$ where n = the number of items on the Grade 2 test.

A total score on each test was thus computed. An example of the cumulative frequency distribution for observed and predicted scores on the level taken is shown in Figure 5.

Traditional norming procedures were then applied to these scores in determining grade equivalents. The median score was found by weighting the contribution of a score of a given student based on the relative proportion of scores obtained from his SES level.

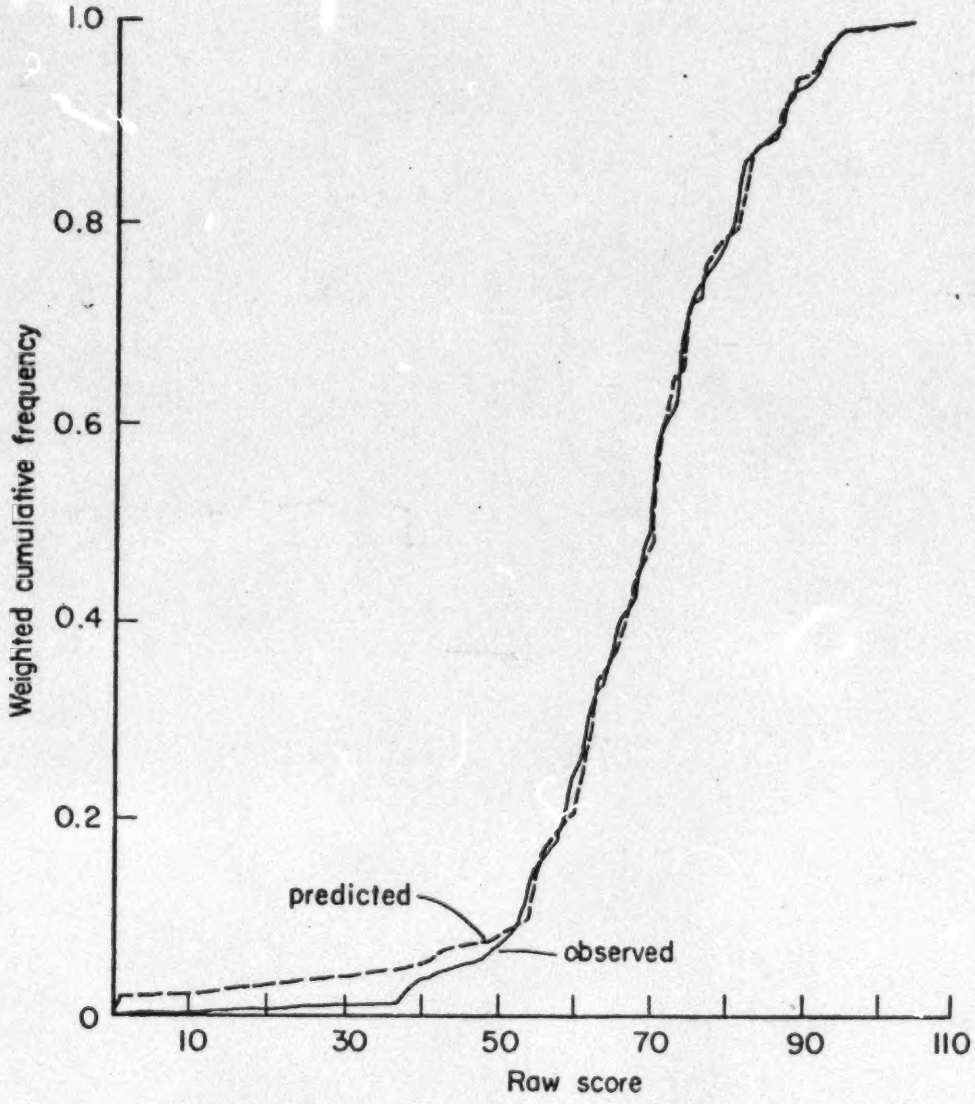


Figure 5. Distribution of observed scores and predicted scores.

A curve was fitted to the data points represented by each grade for a given test. Figure 6 displays a set of such points. Interpolation yielded grade placement by tenths of a grade.

Using the procedures described above, six levels of the mental arithmetic tests for elementary school children were developed and normed.

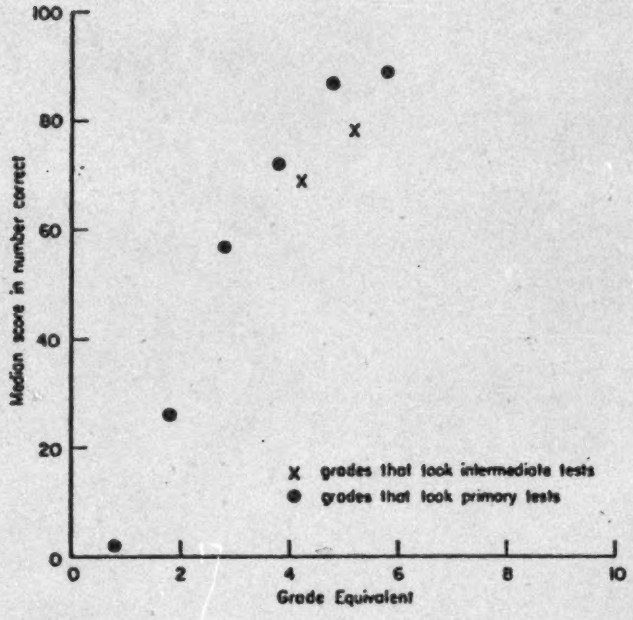


Figure 6. Score by grade equivalent, Primary I.

## REFERENCES

American Psychological Association, French, J. W., & Michael, W. B., cochairman. Standards for educational and psychological tests and manuals. Washington, D. C.: Authors, 1966.

Anscombe, F. J. The transformation of poisson, binomial and negative-binomial data. Biometrika, 1948, 35, 246-254.

Cox, D. R. The analysis of binary data. London: Methuen, 1970.

Olshen, J. S. The use of performance models in establishing norms on a mental arithmetic test. Stanford, CA: Institute for Mathematical Studies in the Social Sciences, 1975. (Technical Report, No. 259)

Suppes, P., Jerman, M., & Brian, D. Computer-assisted instruction: Stanford's 1965-66 arithmetic program. New York: Academic Press, 1968.

Suppes, R., & Morningstar, M. Computer-assisted instruction at Stanford, 1966-68. New York: Academic Press, 1972.

## REFERENCE NOTES

1. Suppes, P., Searle, B., & Lorton, P. The Stanford CAI arithmetic strands program. (In preparation)